Chapter 5 SOLARIA: a protocol for automated cross-peak assignment and structure calculation for solid-state magic-angle spinning NMR

5.1 Introduction

Recently, is has been shown that structures of proteins can be determined by magic-angle-spinning (MAS) solid-state NMR[20] (§ 1.3.1). Central to most structure determination procedures is the collection of a set of distance restraints that is sufficiently large to achieve convergence in the calculations. Such manual assignment step can form a time-limiting factor in the structure determination process. Due to the effects of spin diffusion in MAS NMR, there are large numbers of cross-peaks in spectra recorded with longer mixing times, which can be highly ambiguous and are difficult to assign consistently when a manual approach is followed. Hence, solid-state NMR structure determination could benefit from automation of the cross-peak assignment and structure calculation procedure.

In this Chapter, a strategy for automated protein structure calculation based on solid-state NMR data is presented, making use of concepts known from solution NMR, such as the ARIA protocol for the assignment of ambiguous cross-peaks. For use with solid-state NMR data, the SOLARIA programme[c], a MAS NMR-dedicated version of ARIA, was created. Distance-dependent carbon-carbon correlation spectra in solid-state NMR share in principle similar features than solution NMR NOESY spectra, so that analogous approaches for automated peak assignment can be applied. Hence, SOLARIA and ARIA share the same overall architecture consisting of nine iterations of cross-peak assignment and structure calculations (§ 1.4.2). However, there are also significant differences. The two most noteworthy ones are: a) in solid state NMR, cross-peak volumes are affected by more parameters than in solution NMR (§ 1.3) and a straightforward uniform calibration as applied to solution NMR data (see Equation 1.30) would lead to erroneous estimations of the boundaries. For this reason, a MAS solid-state NMR structure calculation strategy was developed, which neglects cross-peak volumes, using the same lower bound and the same generous upper bound for all constraints. b) The labelling pattern of the solid-state NMR samples is exploited for better convergence of the automated assignment process. A more detailed view of the situation is given in the following section.

5.2 Computational aspects: main differences between SOLARIA and ARIA

SOLARIA borrows the overall iterative architecture, as well as many routines, from ARIA v1.2. In each iteration, peak lists are first annotated (§ 1.4.2.1) and then merged (§ 1.4.2.3) into a single peak list where redundant cross-peaks appearing in different spectra are removed. Subsequently, cross-peaks with several assignment possibilities are transformed into ambiguous distance restraints, which are then used to calculate a family of structures with the simulated annealing protocol CNS. Hereafter, the typical aspects of solid-state NMR that required modifications of the ARIA protocol are discussed § 5.2.1-5.2.5.

[c] available from www.fmp-berlin.de

5.2.1 Different types of correlations observed in distance-dependent solid-state NMR and NOESY-type solution NMR spectra

In liquid-state NMR, the collection of long-range distance restraints for structure calculation protocols relies on the NOE between pairs of protons. In solid-state NMR, carbon-carbon and carbon-nitrogen correlations are more commonly used for extracting distance information, and the cross-peak volumes show either a r^{-3} or r^{-6} dependency. Taking all this into account, SOLARIA allows also to treat distance-dependent correlations between heteronuclei. The user can specify which kind of correlation is measured (homonuclear ^{13}C-^{13}C, ^{15}N-^{15}N or even heteronuclear ^{13}C-^{15}N correlation data).

5.2.2 Carbon-labelling pattern

SOLARIA accounts for carbon-labelling patterns when assigning cross-peaks. The user can indicate the kind of carbon labelling strategy used for each experiment, choosing among the traditional uniform labelling, or the labelling obtained by expression of the protein on a medium containing [1,3-^{13}C]-glycerol or [2-^{13}C]-glycerol as sole carbon sources (§ 1.3.1.1). Other labelling patterns can be easily implemented, for instance for proteins obtained from growth media containing selectively ^{13}C-labelled succinic acid as a precursor[59]. For the glycerol-based strategies, the twenty amino acids can be divided in two groups of ten, labelled A and B, as shown in Figure 1.11. The labelling patterns are taken into account in SOLARIA in the following way:

a) Since non-labelled carbon sites (mostly in amino acids of group A) cannot give rise to cross-peaks, their resonance frequencies must not be used to assign peaks. Depending on which sample was used, the programme automatically deletes these resonances from the list of resonance assignments used for the peak annotation, prior to calculation.

b) The software removes all assignment options involving connected carbons within the same residue that are prohibited according to the alternated labelling. In particular, the [2-^{13}C]-glycerol labelling scheme is predominantly alternating, except for the two cases of valine and leucine. This may not be immediately apparent from the inspection of the labelling of residues in group B in Figure 1.11. The patterns in the figure display an average over several isotopomers and may give the false impression that for these amino acids there are still many cases in which connected nuclei are simultaneously labelled. This is, however, not the case if one considers the labelling schemes of the different isotopomers (see, for example, the labelling of the isotopomers of arginine in Figure 1.11, bottom section. All arginine isotopomers show separated ^{13}C labels when [2-^{13}C]-glycerol is used for labelling.

c) Finally, the percentage of labelling in each carbon position is accounted for in the weighting factor w_{ij} applied to the different assignment options, prior to the structure calculations. This is described in detail hereafter.

5.2.3 Definition of a labelling- and power-adapted weighting factor w_{ij}

In SOLARIA, like in ARIA, during the iterations an increasing fraction of less-representative assignment options is excluded from the summation in Equation 1.29, prior to the structure calculation, by means of the parameter *p* (§ 1.4.2.1.2).

Depending on the solid-state spectroscopy used, SOLARIA chooses the most appropriate distance dependencies to be used as a criterion to weight assignment options according to their distance, as suggested by Equations 1.17 and 1.18. Additionally, a new factor ($s_i{\cdot}s_j$) was introduced into the calculation of the weights, which accounts for the labelling percentages of carbons in labelled samples. Owing to this factor, the relative weight of an assignment option is preferably directly proportional to percentages of labelling of the two interacting nuclei.

Following this discussion, the weighting factors w_{ij} for different assignment options are calculated in SOLARIA by the following modified version of Equation 1.26:

$$w_{ij} = \hat{d}_k^n \cdot s_i s_j , \qquad 5.1$$

with n = -3 or -6, depending on the type of spectroscopy applied.

5.2.4 Fixed boundaries for distance restraints

Due to the difficulties to interpret volumes in terms of distances, fixed values were used for the lower and upper limit of each distance restraint. For the lower limit, a value of 2.8 Å was chosen, which corresponds to the sum of the Van der Waals radii for carbons[38]. For the upper limit, a default value of 6.5 Å was chosen, which is larger than what is commonly used in solution NMR, but smaller than the largest distances observed in solid-state spectra in previous studies[20]. The user can specify this upper limit for each spectrum individually during the setup of the calculation, which is useful if spectra recorded at intermediate mixing times are used.

5.2.5 Inter-molecular peaks

From a computational point of view, inter-molecular cross-peaks (§ 1.3.1.2) in peak lists from PDSD spectra are not different of artefacts or noise peaks, in that they all tend to generate distance restraints which are inconsistent with the structure of the monomer. Hence, a peak list that contains several inter-molecular constraints can be considered as an example of a particularly noisy peak list, but does not represent a new computational challenge. On the other hand, it is important to realise that inter-molecular cross-peaks *do* differ from

„traditional“ artefacts in that they establish systematic, rather than random, correlation networks between residues. As a consequence, they may be more persistent in inducing local distortions in the structures. Conversely, these important characteristics can be exploited to identify them by searching in the list of discarded peaks for correlations rejected systematically.

For this, SOLARIA was programmed to perform a statistic analysis on the rejected peaks at the end of the calculation for detecting inter-molecular contacts. In this routine, the programme simply lists the initial N_δ assignment options for all rejected peaks and tries to identify consistent networks of interactions among them. It is important to mention that in the case that the calculations produce local distortions, such as backbone distortions or wrong orientations of the side-chains, some correct intra-molecular networks of cross-peaks may be also systematically rejected, since they are inconsistent with the distorted structures.

5.3 Results and discussion

SOLARIA was tested on lists with the coordinates of manually picked peaks from PDSD-type spectra of the α-spectrin SH3 domain. Two different sets of peak lists were used: peak lists where inter-molecular cross-peaks were manually removed and more realistic solid-state peak lists where both intra- and inter-molecular cross-peaks were present. In general, when using peak lists obtained from solid-state data, the computational task for automated resonance assignment can be particularly cumbersome. First, generous values for the chemical shift tolerances have to be chosen in order to compensate for the poor resolution of solid-state spectra, resulting in a high number of assignment options per peak (the average number of assignment candidates per peak was approximately 6 for the two 3D spectra, and 19 and 16 for the 1,3-CC and 2-CC 2D spectra, respectively). In addition to this, large boundaries have to be used for all distance restraints (§ 5.2.4). As a consequence, distance restraints in calculations with solid-state data bear not only higher ambiguities, but are also extremely loose.

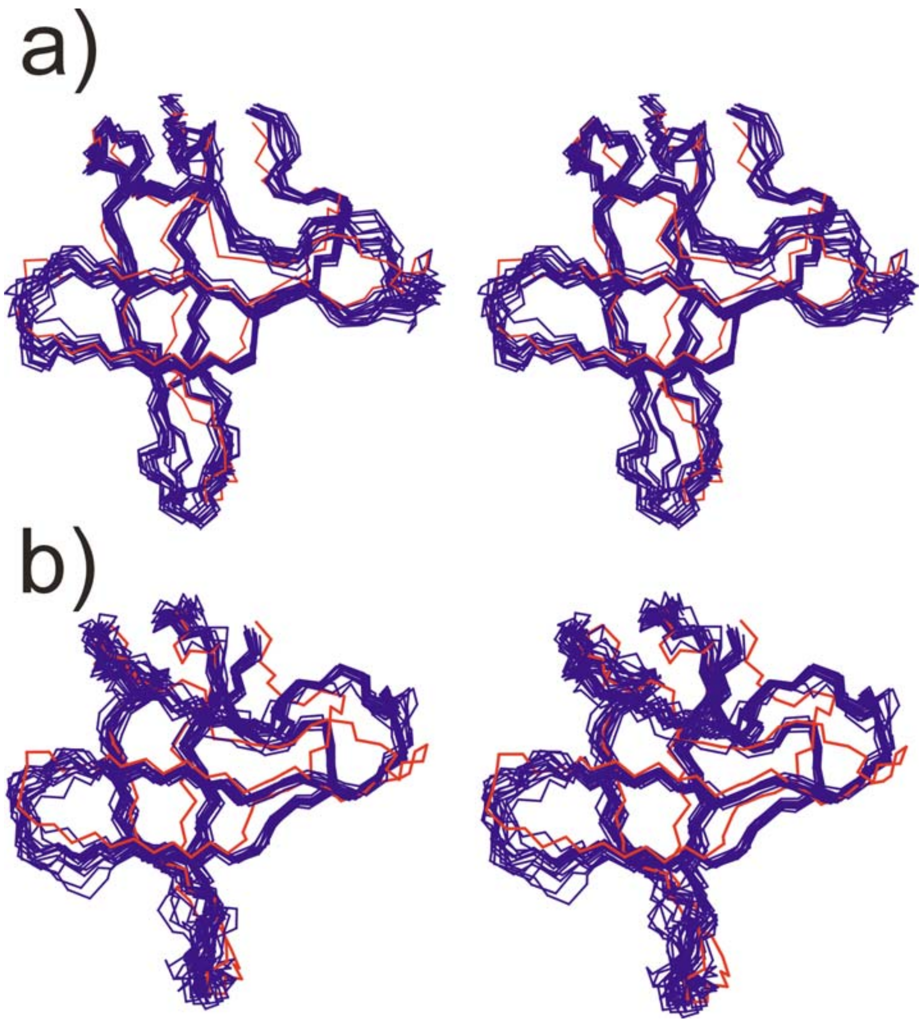

Figure 5.1 Stereo-view of the eleven lowest-energy solid-state NMR structures of the α-spectrin SH3 domain calculated by SOLARIA, coloured in blue. For comparison, the X-ray structure (PDB entry 1SHG) is included, displayed in red and overlaid with the family of solid-state structures by fitting the backbone atoms to the average solid-state structure. a) Calculations were performed on peak lists where inter-molecular cross-peaks were manually removed. b) Calculations were performed on more realistic peak lists containing also inter-molecular cross-peaks.

To face the computational difficulties inherent to the use of solid-state peak lists, the convergence capability of the software was enhanced by substantially increasing the number of cooling steps, as suggested in Chapter 4. The total number of cooling steps was increased from 9,000, which is the ARIA default value, to 100,000. Under these conditions, and using standard ARIA v1.2 input values for all other parameters, SOLARIA produced convergent results for both sets of peak lists. In Figure 5.1a, the results of calculations using lists containing only intra-molecular cross-peaks are presented, while Figure 5.1b shows the results obtained with the second set of lists, containing both intra- and inter-molecular cross-peaks. In both figures, the 11 lowest-energy structures (blue) are overlaid with the X-ray structure (red) as reference (PDB entry: 1SHG). Well-defined structure ensembles with a precision of 0.73 Å and 0.75 Å rmsd, respectively, were obtained despite the use of extremely generous boundaries for distance restraints. When the inter-molecular cross-peaks were excluded, SOLARIA produced accurate structures, with 1.3 Å rmsd to the reference. Less accurate structures (2.2 Å backbone rmsd to the reference) were obtained when the inter-molecular cross-peaks were kept in the peak lists.

Compared to the structures in Figure 5.1a, those in Figure 5.1b are characterised by small local distortions due to an erroneous assignment of a few inter-molecular cross-peaks to intra-molecular contacts. In particular, the regions formed by the residues 35-40 and 54-58 are inaccurate, mainly because ambiguous cross-peaks due to inter-molecular correlations between Ile 30 and Ser 36 are incorrectly assigned to intra-molecular correlations between Ser 36 and Ala 56. Nonetheless, the global fold correctly describes the β-sandwich fold of the SH3 domain.

The surprisingly good results with the peak lists containing also inter-molecular cross-peaks can be attributed to the noise-recognition routine[32] of ARIA, implemented in SOLARIA, which successfully identified and rejected most of the inter-molecular correlations, since they

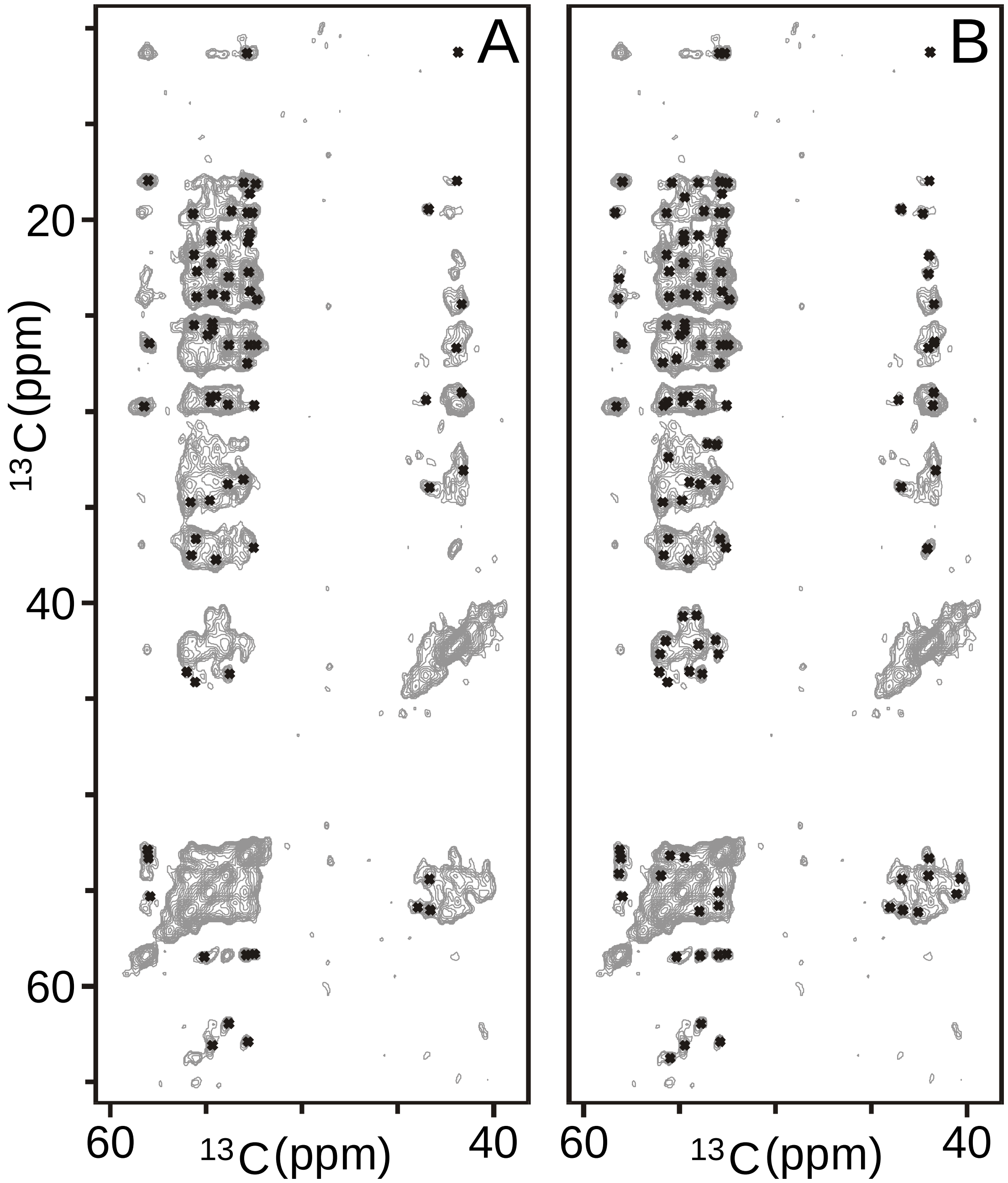

Figure 5.2 Visualisation of the assignments. A strip of a two-dimensional PDSD spectrum recorded using [1,3-^{13}C]-glycerol-grown α-spectrin SH3 domain sample is shown. In (A), the cross-peaks that could be assigned manually using 2D and 3D data are indicated. In (B), cross-peaks assigned and used for the structure calculations by SOLARIA using the same dataset are indicated.

resulted in distance restraints that were inconsistent with the calculated structures (§ 1.4.2.5). The SH3 domain is a small protein with a large surface-to-core ratio. This ratio drops with increasing protein size, hence it is expected that the impact of inter-molecular peaks on the calculation will decrease with larger proteins. In this respect, the small SH3 domain can be considered as a good test case to test SOLARIA's robustness in working with realistic solid-state peak lists containing both intra- and inter-molecular cross-peaks.

Both structures of Figure 5.1a and 5.1b were obtained after approximately 12 hours of calculation time, which is in striking contrast to the several months required for the manual assignment of the same spectra[20]. More importantly, SOLARIA allowed the assignment of approximately 20% more cross-peaks than in the previous manual assignment procedure. In Figure 5.2, all manual assignments obtained from the evaluation of 2D and 3D spectra are indicated in A, displayed in a region of a 2D experiment, and all automatically assigned peaks are indicated in B. The figure clearly indicates that also cross-peaks in highly overlapped regions of the spectrum can be handled and included in the calculation if an automated approach is followed.

As outlined above, inter-molecular cross-peaks occur as consistent networks of signals in the spectrum which tend to be systematically rejected during the calculations. This can be used to identify them. At the end of the calculation, SOLARIA is programmed to search for patterns among the assignment options of the rejected peaks. These patterns are defined by two or more interactions between residues i and $i \pm 1$ and j and $j \pm 1$. In figure 5.3, the interaction networks identified by SOLARIA are displayed in a 2D correlation grid. It contains the inter-molecular contacts but also systematically rejected intra-molecular contacts in inaccurately determined regions of the structure. However, if convergence was not too poor, the latter should represent only a small percentage of the total number of the networks present in the grid. In both cases, the networks identified by the analysis can either be unambiguous or ambiguous. In the latter case the network originates from ambiguously rejected cross-peaks that have more than one assignment option in common.

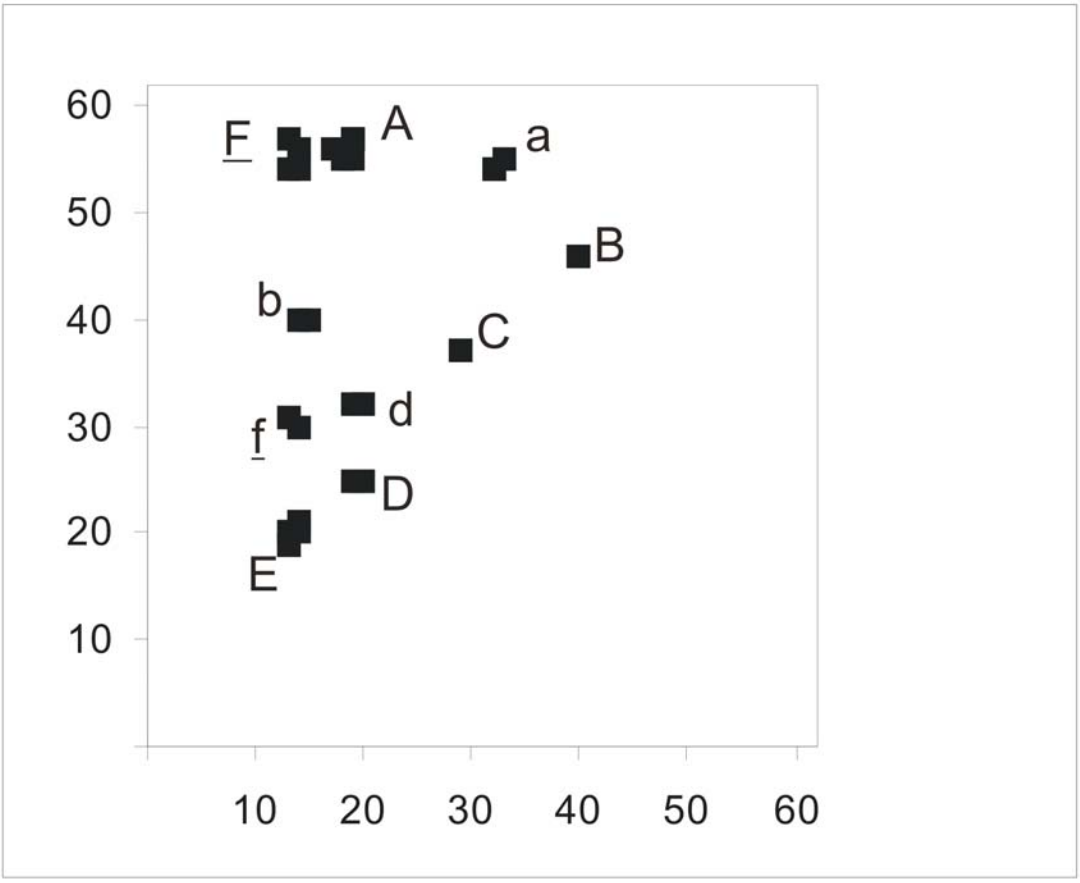

Figure 5.3 Networks of inter-residual contacts identified by SOLARIA among the assignment options of the peaks rejected during the calculation. The selected networks are displayed in a 2D plot; on both axes, the residue numbers for the SH3 domain are shown. The networks can be either unambiguous (letters C and E) or ambiguous. The latter originate from the rejection of cross-peaks displaying more than one assignment option in common. Groups of contacts within the same ambiguous network are labelled with the same letter. The group of contacts which has physical meaning is indicated by a capital letter, while assignment options that have no physical meaning are shown with small letters (A-a, B-b, D-d, F-f). Five networks out of six (letters A-a, B-b, C, D-d, E) were generated by inter-molecular contacts, whereby one network was due to a systematic rejection of intra-molecular contacts within the distorted region 54-58 of the structure (underlined letters F and f).

These ambiguous networks are displayed as two groups of contacts labelled with the same letter in Figure 5.3. However, only one group of assignment options has a physical meaning, and it is labelled with a capital letter; assignment options that have no physical meaning are shown with small letters.

In order to verify the reliability of the analysis, the X-ray structure was used to assess which contacts in the grid correspond to inter-molecular interactions. Five out of the six networks correctly identify an inter-molecular contact, whereas only one represents rejected intra-molecular cross-peaks from an incorrectly determined region of the calculated structures (Figure 5.4).

The latter is indicated with underlined letters in Figure 5.3. Hence, the networks selected by the analysis are mainly due to inter-molecular contacts, and, to a minor extent, to a set of intra-molecular contacts within an inaccurately determined region of the structure. Such an analysis could find applications during an automated refinement of the surface regions of the structures, as outlined in the next paragraph.

5.4 Conclusions and perspectives

The software SOLARIA was used to assign in a completely automated manner cross-peaks from 2D and 3D PDSD solid-state spectra of the α-spectrin SH3 domain. To achieve this, the solution NMR-oriented ARIA programme was partly rewritten, and new routines to account for many aspects typical of solid-state NMR were introduced. The programme accepts peak lists containing not only ^{1}H-^{1}H, but also typical solid-state ^{13}C-^{13}C, ^{15}N-^{15}N and ^{13}C-^{15}N correlations. The reduced labelling obtained when [1,3-^{13}C]- and [2-^{13}C]-glycerol are used for preparing the samples is taken into account to simplify the assignment of the cross-peaks. To account for the poor dependence of the peak volumes on the distance, the programme allows for the use of constant boundaries for distance restraints. Hence, input peak lists do not require the presence of cross-peak volumes or intensities. The present version of SOLARIA is not limited to the use of PDSD spectra. In principle, other solid-state NMR pulse sequences, with their specific distance-dependence of cross-peak intensities, can be easily implemented.

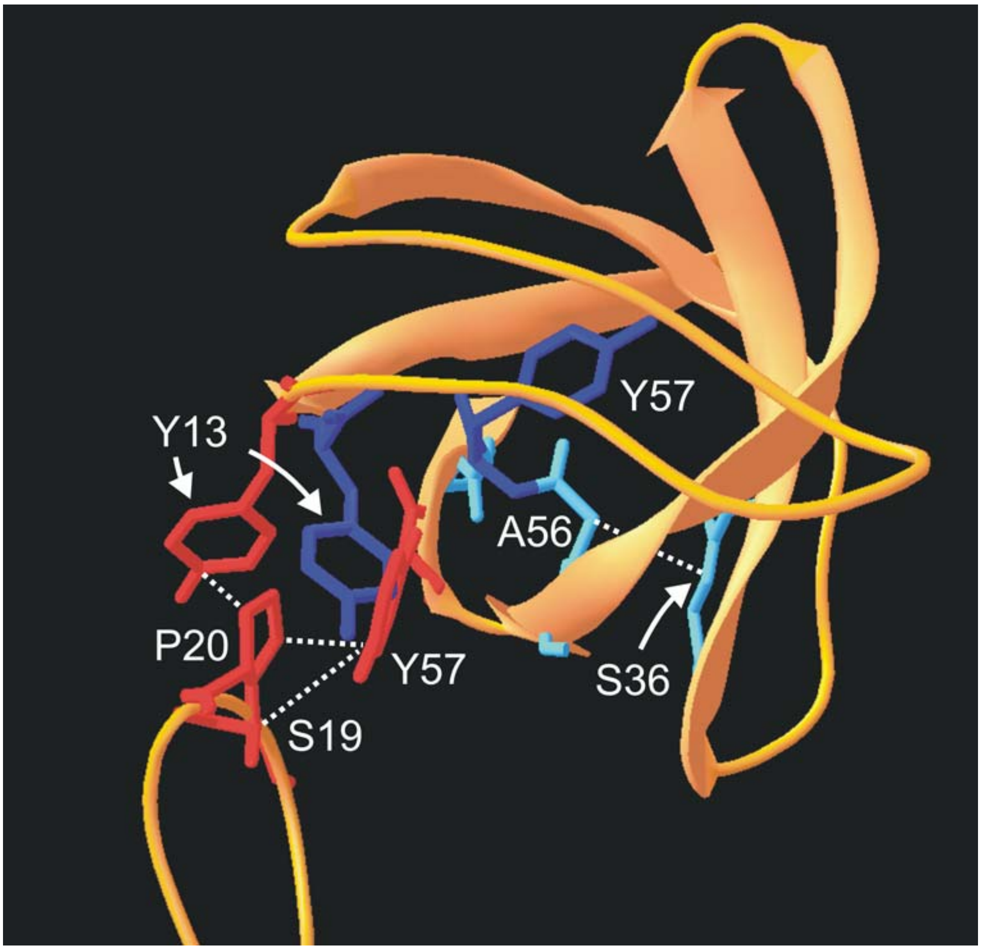

Figure 5.4 The distorted region (54-58) of the backbone of the structures of Figure 5.1b is indicated in light-blue on top of the X-ray reference structure (orange). Sidechains of residues Tyr 13 and Tyr 57 are shown, both for the calculated structure (dark blue) and for the X-ray reference (red). In addition, the residues Pro 20 and Tyr 19 from an adjacent monomer, giving rise to inter-molecular contacts with both Tyr 13 and Tyr 57, are shown in red. Cross-peaks due to these inter-molecular interactions originate restraints which cannot be accommodated in the monomer and are rejected (letter E in Figure 5.3). As a result, both Tyr 13 and Tyr 57 are ‘free’ to assume an incorrect orientation in the calculated structure. In particular, Tyr 57 points towards the core instead of protruding from the surface. This arises because the correct inter-molecular correlation between Ile 30 and Ser 36 is incorrectly assigned to an intra-molecular correlation between Ser 36 and Ala 56. The separation between the two tyrosine rings is now larger than in the correct structure, leading to the rejection of the network of intra-molecular restraints between them (indicated with the letter F in Figure 5.3).

When inter-molecular cross-peaks were manually excluded from the calculation, the calculated structures showed a 1.3 Å rmsd to the X-ray reference. These structures exemplify a first successful attempt to introduce automation to structure determination of solid proteins by solid-state NMR. As a result, the time necessary to evaluate distance-restraints by assigning cross-peaks could be reduced from several months of painstaking manual work of an expert spectroscopist to few hours of calculation time. Most importantly, even ambiguous cross-peaks from solid-state spectra can now be handled and directly included in the calculations. Furthermore, by analysing the inconsistent cross-peaks rejected by the software, an automated strategy was set up to identify most of the regions of inter-molecular contact within the solid.

These results show that the presence of inter-molecular peaks in the peak lists does not constitute a serious problem for convergence and that SOLARIA can be used to obtain a good starting structure for the monomer from completely unassigned and unfiltered PDSD peak lists. Assuming that the symmetry of the solid packing is known (e.g. from EM measurements), this starting structure could in turn be used to build an approximate template of the repetitive unit of the solid and gain knowledge on the inter-molecular regions. A sufficiently realistic initial template structure of the repetitive unit of the solid would represent a tremendous help for convergence and would finally provide internal criteria to assign inter-molecular cross-peaks. The structure of the monomer together with the regions of contact between different monomers would then be refined simultaneously in an iterative fashion by assignment of all entries of the peak list, including the inter-molecular cross-peaks. Hence, in the near future, SOLARIA might open the way to automated solid-state NMR structural investigations of biologically relevant systems like amyloidal fibrils or virus capsids.